British Journal of Industrial Medicine 1985;**42**:346–349

Some factors influencing interobserver variation in classifying simple pneumoconiosis

D C MUSCH,[1] I T T HIGGINS,[1] AND J R LANDIS[2]

From the Departments of Epidemiology[1] and Biostatistics,[2] School of Public Health, University of Michigan, Ann Arbor, Michigan 48109, USA

ABSTRACT Three experienced physician readers assessed the chest radiographs of 743 men from a coal mining community in West Virginia for the signs of simple pneumoconiosis, using the ILO U/C 1971 Classification of Radiographs of the Pneumoconioses. The number of films categorised by each reader as showing evidence of simple pneumoconiosis varied from 63 (8·5%) to 114 (15·3%) of the 743 films classified. The effect of film quality and obesity on interobserver agreement was assessed by use of kappa-type analytic procedures for measuring agreement on categorical data. Poor film quality and obesity both affected agreement adversely. Poor quality films were disproportionately frequent in obese individuals, as defined by the Quetelet index. On control of film quality by stratification, the effect of obesity on interobserver profusion agreement was no longer evident.

Disagreement between observers on the interpretation of chest radiographs for the pneumoconioses has been reported.[1–3] The United States Federal Mine Safety and Health Act of 1977 (PL-95-164) called for miners with "evidence of pneumoconiosis" to be transferred, without loss of pay, to areas where dust concentrations are maintained at less than 1 mg/m^3 of respirable dust. Of the three means of providing evidence of pneumoconiosis (chest radiography, biopsy, or necropsy), the chest roentgenogram is of primary importance. Thus it is clearly important that the radiographic evidence of pneumoconiosis is reliably interpreted.

One factor that would seem to interfere with consistent radiographic interpretation is poor film quality. When Liddell compared discrepancies among six readers on the best and worst quality radiographs from 11 coal miners, he concluded that poor film quality did not significantly affect variability in film reading.[4] Several studies by Reger and his colleagues addressed the impact of film quality on consistency of interpretation. In the first the authors evaluated the results of three experienced radiologists' interpretations of 2337 radiographs and found only a trivial effect of poor quality on observer consistency.[2] In the second interobserver consistency was adversely affected by poor film quality (especially by underpenetrated films).[5] Finally, Musch *et al* found a significant adverse effect of poor film quality on agreement between three readers in classifying chest radiographs from 1771 taconite workers for pneumoconiotic changes.[6]

There is some evidence that radiographs from obese individuals are more likely to be poor in quality. One study found an indirect relation between chest thickness (measured at maximum inspiration) and radiographic quality.[4] The author concluded that "there are some subjects, especially those with thick chests, in whom it is difficult to get technically good films, and it is in these that the effects of poor technique probably affect the accuracy of reading of pneumoconiosis most." Another study related the proportion of unsatisfactory quality films to the subjects' weight/sitting height (W/H) ratio, and found this proportion to increase with increasing W/H ratios.[7] This trend held true for three separate measures of technical quality (density, contrast, and definition). It is the purpose of this investigation to delineate the impact of film quality and obesity on interobserver agreement in classifying radiographs for the profusion of small opacities seen in pneumoconiosis.

Received 30 July 1984
Accepted 17 September 1984

Materials and methods

To ascertain the prevalence of pneumoconiosis in residents of Mullens, a small mining town in West

Virginia, three experienced physicians (NIOSH-certified B-readers) independently interpreted the chest radiographs of 392 coal miners and 351 non-miners. The readers used the extended, 12-point scale of the profusion classification of the ILO U/C 1971 International Classification of Radiographs of the Pneumoconioses.[8] For reference during reading, they used the standard radiographs for the classification scheme. The readers assessed overall radiographic quality by use of the following scale: (1) good, (2) fair, (3) poor, and (4) unreadable. Obesity was assessed by means of the Quetelet index (w/h^2). Since we wished to evaluate interobserver agreement on major categories of profusion of combined small opacities (rounded or irregular opacities less than 1 cm in diameter or width), on completion of the readings we combined the subcategories of profusion into major categories as follows:

Major category	Subcategory
Category 0	0/-, 0/0, 0/1
Category 1	1/0, 1/1, 1/2
Category 2	2/1, 2/2, 2/3
Category 3	3/2, 3/3, 3/4

Interobserver variation was analysed using the method for the analysis of multivariate categorical data described in Landis and Koch,[9 10] using the supporting programming for this analysis.[11] The Kappa type measure of agreement used in this analysis summarises the standardised interobserver agreement on profusion (number per unit area) of simple opacities, classified as categories 0, 1, and 2 and 3. When the degree of abnormality in a population is low, two (or more) readers would be expected to agree on most classifications by chance alone. This standardised agreement index (κ) corrects observed agreement among readers for that expected by chance, and is commonly defined as: $\kappa = (\pi_o - \pi_e)/(1 - \pi_e)$, where π_o = observed agreement and π_e = expected agreement. With complete agreement, $\kappa = +1$; values between 0 and +1 indicate agreement greater than or equal to chance agreement; values less than 0 indicate agreement less than chance would predict. Contrasts of kappa values are discussed in Landis and Koch[9 10] and may be readily computed within the scope of the computing framework outlined in Landis, Stanish, and Koch.[11]

Results

The three readers classified from 8·5% to 15·3% of the radiographs as showing evidence of simple pneumoconiosis (table 1). Most of the differences between readers occurred in the use of profusion categories 0 and 1. Interobserver pairwise agreement (agreement between reader pairs on individual radiographs) shows that readers B and C agreed more closely than did either with reader A (table 2). On comparing Kappa values, this difference was statistically significant ($p < 0·0001$). All three readers coincided on the profusion classification (major category) of 87·8% (652) of the radiographs. When this observed agreement is adjusted for that expected by chance alone, the resultant kappa value is 0·60 (SE = 0·04).

To evaluate the effect of film quality on interobserver profusion agreement, use of the film quality scale by the readers was examined. As table 3 shows, no two readers used the scale in a similar manner. While none of the readers used the "unreadable" category, most of the disagreement between readers occurred on distinguishing "good" from "fair" quality. Therefore, these two categories were combined, and the radiographs were separated into two groups based on defining satisfactory quality radiographs as those which received no rating less than "fair" from the three readers. Of the 743 radiographs, 91·9% (683) met this criterion. An unsatisfactory quality radiograph, then, has a rating of "poor" quality by at least one of the readers.

Contrast of interobserver joint agreement—that

Table 1 *Profusion classifications by reader*

Reader	*Major profusion category*		
	0 %(n)	*1* %(n)	*2 & 3* %(n)
A	84·7 (629)	8·7 (65)	6·6 (49)
B	89·8 (667)	5·1 (38)	5·1 (38)
C	91·5 (680)	3·2 (24)	5·3 (39)

n = 743.

Table 2 *Interobserver pairwise profusion agreement*

Reader pair	*Obs*	*Exp*	$\hat{\kappa}$	*SE*
A and B	0·8896	0·7678	0·5245	0·0419
A and C	0·9058	0·7811	0·5697	0·0431
B and C	0·9556	0·8259	0·7450	0·0396

n = 743.

Table 3 *Film quality classifications by reader*

Reader	*Film quality category*		
	Good %(n)	*Fair* %(n)	*Poor* %(n)
A	77·4 (575)	19·5 (145)	0·1 (1)
B	7·5 (56)	89·4 (643)	3·1 (22)
C	40·9 (303)	54·1 (383)	5·0 (35)

n = 743.

is, agreement between all three readers—on the profusion of small opacities between radiographs of satisfactory and unsatisfactory quality shows the kappa statistic to be significantly greater ($p < 0.01$) for radiographs of satisfactory quality (table 4). If at least one of the readers classified the radiograph quality as "poor," agreement on the interpretation of profusion on those radiographs is adversely affected. This finding also holds for each reader pair.

The effect of obesity on consistency of profusion categorisation was evaluated by contrasting interobserver agreement on radiographs from subjects in the lower four and upper quintiles of the sample's Quetelet index distribution (using the 80th percentile, 29·7 Kg/m², as an arbitrary cut point). The upper quintile would be expected to contain most of the obese (as well as heavily muscular) members of the Mullens sample. The mean Quetelet index value for the lower four quintiles was 25·2 Kg/m²; the corresponding value for the upper quintile was 32·8 Kg/m². This latter value would be equivalent to a relative weight about 140% and 130% of the "desirable" weight for men of medium and large frame sizes, respectively, based on the Metropolitan Life Insurance tables.[12] A statistically significant difference ($p < 0.01$) is shown in interobserver joint profusion agreement on radiographs from the two groups defined in terms of their Quetelet indexes (table 5). A similar difference was seen for each pair of readers.

Since the adverse effect of obesity on interobserver profusion agreement may be a result of confounding by film quality, this possibility was evaluated. For the individuals in the upper quintile of the Quetelet distribution, 12·3% had unsatisfactory quality radiographs. The corresponding figure for those in the lower four quintiles was 6·9%. This difference exceeds that expected by chance ($\chi^2 = 3.89$; $p < 0.05$). The effect of obesity, then, could be due to poorer film quality in obese individuals' radiographs. On control of film quality by stratification, there is no significant effect of obesity on interobserver agreement (table 6).

Discussion

We found substantial differences between the three readers in their classification of the profusion of combined, small opacities on 743 radiographs. If we define evidence of pneumoconiotic changes on the radiographs as profusion category 1, 2, or 3 the prevalence of these changes varied from 8·5% (reader C) to 15·3% (reader A). While this almost twofold range is less than the over threefold range (7·0–23·0%) of pneumoconiosis prevalence determined by three readers in another study,[6] it provides clear evidence of interobserver variation in classifying radiographs for pneumoconiosis. Such variation would be especially disconcerting to workers seeking medical evidence to substantiate a request for transfer or disability compensation.

Most of the discrepancies between readers occurred in discriminating between major profusion categories 0 and 1. Category 0 refers to the absence

Table 4 *Interobserver joint profusion agreement by film quality*

Film quality	*No*	*Obs*	*Exp*	$\hat{\kappa}$	*SE*
Satisfactory	683	0·9019	0·7297	0·6370*	0·0378
Unsatisfactory	60	0·6000	0·3687	0·3664	0·0909

*$p < 0.01$, for the comparison of Kappa values from the readers' assessment of satisfactory and unsatisfactory quality radiographs.

Table 5 *Interobserver joint profusion agreement by Quetelet index level*

Quetelet index percentile	*No*†	*Obs*	*Exp*	$\hat{\kappa}$	*SE*
≤80	590	0·9136	0·6919	0·7194*	0·0381
>80	146	0·8425	0·7417	0·3902	0·1046

*$p < 0.01$, for the comparison of Kappa values from readers' assessment of radiographs of subjects below and above 80th percentile of Quetel index distribution.
†Information on weight or height was missing on seven subjects.

Table 6 *Interobserver joint profusion agreement by Quetelet index level on radiographs of satisfactory quality*

Quetelet index percentile	*No*	*Obs*	*Exp*	$\hat{\kappa}$	*SE*
≤80	549	0·9271	0·7204	0·7394*	0·0410
>80	128	0·9141	0·7893	0·5921	0·1110

*$p > 0.20$, for comparison of Kappa values from readers' assessment of satisfactory quality radiographs of subjects below and above 80th percentile of Quetelet index distribution.

of opacities or presence of opacities less profuse than category 1, whereas category 1 shows small opacities present but few in number. Even though standard, mid-category radiographs were available for reference, the readers had the most difficulty in deciding on the proper profusion category for radiographs showing minimal changes that may or may not be due to pneumoconiosis. When faced with distinguishing between category 0, which includes opacities present but less profuse than category 1, and category 1, which is reserved for opacities that are few in number, the reader's decision is influenced by past training, experience, and attitudes toward minimal radiographic change.

The finding that obese individuals' radiographs had a greater proportion of poor quality films is consistent with information on causes of poor radiographic quality. Density and contrast, two important factors that influence radiographic quality, are both affected by the thickness of tissue through which the x ray must pass.[13] On analysing interobserver agreement on good quality radiographs from subjects in the lower four and upper Quetelet index quintiles (table 6), the significant effect of obesity evident in table 5 is no longer present. Poor film quality, then, is the cause of reduced interobserver agreement on obese individuals' radiographs.

Our study provides support for other studies[5 6] that have found an adverse effect of poor film quality on the consistency of interpretation by multiple readers. It is important to note that factors other than film quality contribute to observer variation. These include conditions of the reading place (lighting, surroundings), observer fatigue, and subjective factors such as differing perceptions of abnormality/normality. Whereas some interobserver variation is to be expected, it is important to identify factors that both influence reliability and are amenable to control. Film quality, we contend, is such a factor.

We thank Drs L Bristol and J Rosentein who, with Dr I T T Higgins, provided the radiological interpretations used in this study. We also thank Mary S Oh, who edited and arranged the computer files for analysis.

Supported in part by Grant No 5-T32-HL07337-03 from the National Heart, Lung, and Blood Institute, National Institute of Health, Bethesda, Maryland, USA.

References

[1] Fletcher CM, Oldham PD. The problems of consistent radiological diagnosis in coalminers' pneumoconiosis. *Br J Ind Med* 1949;**6**:168–83.

[2] Reger RB, Morgan WKC. On the factors influencing consistency in the radiologic diagnosis of pneumoconiosis. *Am Rev Respir Dis* 1970;**102**:905–15.

[3] Felson B, Morgan WKC, Bristol LJ, *et al.* Observations on the results of multiple readings of chest films in coalminers' pneumoconiosis. *Radiology* 1973;**109**:19–23.

[4] Liddell FDK. The effect of film quality on reading radiographs of simple pneumoconiosis in a trial of x-ray sets. *Br J Ind Med* 1961;**18**:165–74.

[5] Reger RB, Smith CA, Kibelstis JA, Morgan WKC. The effect of film quality and other factors on the roentgenographic categorization of coal workers' pneumoconiosis. *American Journal of Roentgenology, Radium Therapy and Nuclear Medicine* 1972;**115**:462–72.

[6] Musch DC, Landis JR, Higgins ITT, Gilson JC, Jones RN. An application of Kappa-type analyses to interobserver variation in classifying chest radiographs for pneumoconiosis. *Statistics in Medicine* 1984;**3**:73–83.

[7] Pearson NG, Ashford JR, Morgan DC, Pasqual RSH, Rae S. Effect of quality of chest radiographs on the categorization of coalworkers' pneumoconiosis. *Br J Ind Med* 1965;**22**:81–92.

[8] Jacobson G, Lainhart WS, eds. ILO U/C 1971 international classification of radiographs of the pneumoconioses. *Med Radiogr Photogr* 1972;**48**:65–76.

[9] Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977;**33**:159–74.

[10] Landis JR, Koch GG. An application of hierarchical Kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 1977;**33**:363–74.

[11] Landis JR, Stanish WM, Koch GG. A computer program for the generalized chi-square analysis of categorical data using weighted least squares to compute Wald statistics (GENCAT). *Comput Programs Biomed* 1976;**6**:196–231.

[12] Metropolitan Life Insurance Company. New weight standards for men and women. *Statistical Bulletin* 1959;**40**:1–4.

[13] Selman J. *The fundamentals of x-ray and radium physics.* 6th ed. Springfield: Charles C Thomas Co, 1977.